4П

УДК 004.934

***В.В. Пилипенко***
Международный научно-учебный центр информационных технологий и систем
г. Киев, Украина
valery_pylypenko@mail.ru

# Распознавание ключевых слов в потоке речи при помощи фонетического стенографа

В статье рассматривается использование фонетического стенографа для распознавания ключевых слов в потоке речи. Для моделирования фонем используются скрытые Марковские модели. Ключевое слово задается последовательностью фонем в виде транскрипции слова. Приведены результаты поиска ключевых слов в потоке речи большого количества дикторов. Предложенный подход может использоваться для поиска речевой информации в огромных массивах данных.

## Введение

В связи с все более активным использованием естественного интерфейса и в частности голоса, для общения с техникой возросло и значение аудиозаписи как носителя информации. Появилась потребность в системах, способных быстро и эффективно обслуживать аудиоархивы и находить нужную информацию в большом объеме записи. Для этой цели предложено использовать алгоритмы поиска ключевых слов в потоке речи.

Задачей поиска ключевых слов является нахождение заданных фрагментов (это могут быть отдельные слова или целые фразы) в потоке речи. Первоначально для задания фрагментов использовались отрезки произнесенной речи, при этом по нескольким произнесениям формировался эталон ключевого слова. Неудобство такого метода проявлялось в том, что для введения в систему нового ключевого слова необходимо заранее его произнести или вырезать из известного потока речи.

Современные алгоритмы поиска ключевых слов используют задание ключевых слов последовательностью фонем или других элементарных единиц. При этом может использоваться преобразователь графема-фонема в соответствии с правилами данного языка и тогда ключевое слово задается текстом слова или фразы, что значительно расширяет область применения такой системы.

Широкое применение получили алгоритмы, в которых для моделирования элементарных единиц уровня фонемы применяются скрытые Марковские модели (СММ). Для поиска ключевых слов используются те же подходы, что и для распознавания слитной речи.

Модификация касается способа задания слов, отсутствующих в словаре системы. Предложено два способа задания неизвестных слов:

1. Моделирование незнакомых слов произвольными последовательностями фонем.
2. Использование Гауссовской Смеси Моделей (*Gaussian Mixture Model GMM)* для моделирования фонового потока речи.

В данной статье рассматривается первый способ задания незнакомых слов. Для этого используется концепция фонетического стенографа [1], [2].

# 1. Базовая система распознавания слитной речи

В данной работе используется инструментарий HTK [3] на основе скрытых Марковских моделей (СММ). При помощи инструментария HTK построены акустические и лингвистические модели системы. Для распознавания речи был разработан программный комплекс, совместимый с акустическими и лингвистическими моделями HTK.

## 1.1. Предварительная обработка речевого сигнала

Речевой сигнал преобразуется в последовательность векторов признаков с интервалом анализа 25 мс и шагом анализа 10 мс. Вначале речевой сигнал фильтруется фильтром высоких частот с характеристикой P(z) = 1-0.97 z-1. Затем применяется окно Хэмминга и вычисляется быстрое преобразование Фурье. Спектральные коэффициенты усредняются с использованием 26 треугольных окон, расположенных в мел-шкале, и вычисляются 12 кепстральных коэффициентов.

Логарифм энергии добавляется в качестве 13-го коэффициента. Эти 13 коэффициентов расширяются до 39-мерного вектора параметров путем дописывания первой и второй разностей от коэффициентов, соседних по времени. Для учета влияния канала применяется вычитание среднего кепстра.

## 1.2. Акустическая модель

В качестве акустических моделей используются скрытые Марковские модели. 56 украинских контекстно-независимых фонем моделируются тремя состояниями Марковской цепи без пропуска. Используется диагональный вид Гауссовских функций плотности вероятности.

Редко встречающиеся фонемы моделируются 64 смесями Гауссовских функций плотности вероятности, более часто встречающиеся фонемы моделируются большим числом смесей, наиболее часто встречающиеся фонемы используют 1024 смеси.

Словарь транскрипций создается автоматически из орфографического словаря с использованием контекстно-независимых правил.

# 2. Акустическое и текстовое наполнение

## 2.1. Обучающая выборка

Обучение производилось на выступлениях депутатов Верховной Рады Украины, записанных через телевизионную сеть. Парламентская речь характеризуется некоторыми особенностями:

– это спонтанная речь. Встречаются отдельные доклады, зачитываемые по подготовленному заранее тексту, однако мало дикторов в точности придерживается этого текста;

– из-за ограничения во времени выступления многих дикторов произносятся в слишком быстром темпе.

Для обучения использовались записи длиной в 197 тыс. секунд, в которых встретилось около 427 тыс. слов. Всего было записано 287 дикторов.

Обучение производилось на предварительно размеченной выборке. Для этого запись выступления автоматически разбивалась на фразы из нескольких слов, ограниченные паузами больше 400 мсек. Среднее количество слов в одной фразе оказалось равным 5.

Каждой фразе оператором ставилась в соответствие метка в виде текста из стенограммы. Затем автоматически производилось преобразование текста в последовательность фонем в соответствии с контекстно-независимыми правилами украинского языка. Выборка, размеченная таким образом, использовалась для построения акустической модели.

## 2.2. Контрольная выборка

Распознавание производилось на выступлениях депутатов, записанных в отличные от обучающей выборки дни. Для распознавания использовались записи длиной в 42 тыс. секунд, в которых встретилось 94 тыс. слов. Всего использовались записи 152 дикторов. Записи 41 диктора не встретились в обучающей выборке. Таким образом, эти дикторы оказались неизвестными для системы распознавания.

## 2.3. Текстовый материал

Словарь был составлен из текстов стенограмм заседаний Верховной Рады Украины. С официального сайта Верховной Рады были загружены все стенограммы заседаний, начиная с 1991 года, что составило больше 100 МБ текста. Текст был модифицирован для того, чтобы убрать служебную информацию из стенограмм (например, аплодисменты), записать числа в текстовом виде, а также отделить русский текст от украинского.

# 3. Фонетический стенограф

Алгоритм фонетического стенографа позволяет строить последовательность фонем для речевого сигнала без использования какого-либо словаря. Для этой цели строится некоторая генеративная грамматика, которая может синтезировать все возможные модельные сигналы непрерывной речи для любой последовательности фонем. В рамках построенной модели строится алгоритм пофонемного распознавания для неизвестного сигнала. Используются те же контекстно-независимые модели фонем, как и в базовом распознавателе.

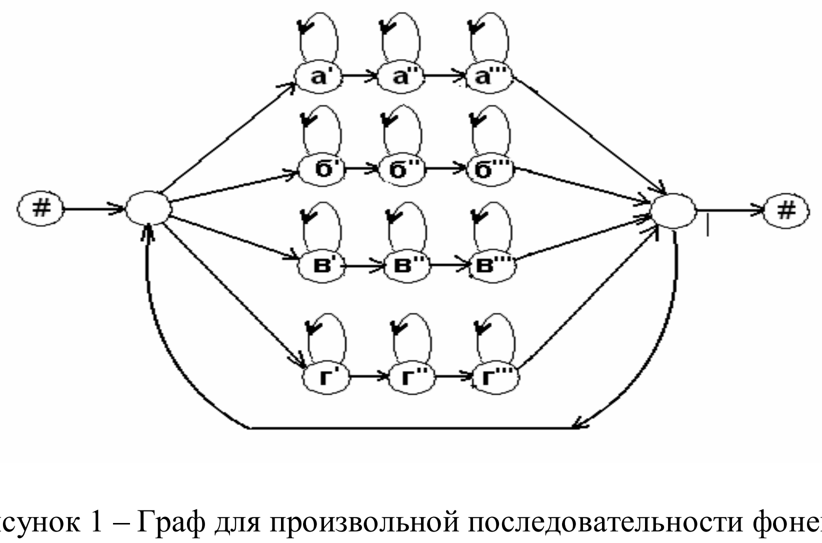



Рисунок 1 – Граф для произвольной последовательности фонем

Надежность найти фонему на правильном месте для известной реализации равна приблизительно 85%.

# 4. Результаты экспериментов по распознаванию ключевых слов в потоке слитной речи

Эксперименты проводились на описанной контрольной выборке.

Ключевые слова описывались последовательностью фонем заданной длины от 2 до 12 фонем. Для данной длины из словаря выбиралось 30 ключевых слов. К сожалению, для длин 2, 11 и 12 в тестовом корпусе не удалось выбрать достаточное количество записей, и в данном случае было выбрано около 20 ключевых слов. Всего было отобрано 309 ключевых слов.

Для каждого ключевого слова из тестового корпуса выбиралось от 15 до 100 записей фраз, в которые это ключевое слово обязательно входило. На данном материале подсчитывался процент *ложного отказа* (False Rejection) как доля случаев, когда ключевое слово не было распознано.

Кроме этого выбиралась выборка длиной в 1000 слов, в которую ключевое слово гарантированно не входило. На данном материале подсчитывался процент *ложного срабатывания* (False Alarm) как доля случаев, когда происходило срабатывание алгоритма распознавания ключевого слова.

Алгоритм содержит коэффициент, позволяющий регулировать соотношение между процентами *ложного отказа* и *ложного срабатывания.* Оптимальный коэффициент был выбран из условия минимума суммы этих процентов. При необходимости можно выбрать другое значение коэффициента, отдавая предпочтение тому или иному сценарию использования системы.

Таблица 1 – Надежность распознавания ключевых слов

| Число фонем в ключевом слове | Процент ложного отказа | Процент ложного срабатывания |
|---|---|---|
| 2 | 6.95 | 13.27 |
| 3 | 5.22 | 7.30 |
| 4 | 3.26 | 4.76 |
| 5 | 4.06 | 2.34 |
| 6 | 3.32 | 1.87 |
| 7 | 2.21 | 1.12 |
| 8 | 1.52 | 1.48 |
| 9 | 2.09 | 0.74 |
| 10 | 3.79 | 0.55 |
| 11 | 4.47 | 0.38 |
| 12 | 5.73 | 0.22 |
| По всем длинам | 3.67 | 3.02 |

В табл. 1 приведены результаты распознавания ключевых слов в зависимости от количества фонем в ключевом слове.

Оптимальное значение коэффициента зависит от длины слова, для более длинных слов его можно увеличить для получения лучших результатов.

# Заключение

Статья описывает экспериментальную систему распознавания ключевых слов в потоке речи на основе фонетического стенографа. Проведены эксперименты по распознаванию. Коэффициент *ложного отказа* равен 3.67% при *ложном срабатывании*, равном 3.02%. Это позволяет надеяться, что данный алгоритм можно использовать в практических системах.

В дальнейшем предполагается рассмотреть комбинацию фонетического стенографа и модели фоновых слов в виде Гауссовской смеси моделей (*Gaussian Mixture Model GMM).*

# Литература

1. Vintsiuk Taras K. Generalized Automatic Phonetic Transcribing of Speech Signals / Taras K. Vintsiuk // Труды Пятой Всеукраинской международной конференции «Оброблення сигналів і зображень та розпізнавання образів» / УАсОІРО. – Київ, 2000. – С. 95-98.
2. Пилипенко В.В. Використання фонетичного стенографа при розпізнаванні мовлення з великих словників / В.В. Пилипенко // Тезисы 12-й международной конференции «Автоматика – 2005». – Харьков, 2005. – С. 73.
3. The HTK Book / [S. Young, G. Evermann, D. Kershaw and others]. – Cambridge University Engineering Department, 2002.

***В.В. Пилипенко***

**Розпізнавання ключових слів у потоці зв'язного мовлення за допомогою фонетичного стенографа**

У статті розглядається фонетичний стенограф для пошуку ключових слів у потоці зв'язного мовлення. Приховані Марківські моделі використовуються для моделювання фонем. Ключове слово задається за допомогою фонетичної транскрипції. Доводяться результати експериментів пошуку ключових слів у потоці мовлення від великої кількості дикторів. Запропонований підхід може використовуватися для пошуку мовної інформації у величезних масивах даних.

***V.V. Pylypenko***

**Keyword Spotting with Using the Phoneme Recognizer**

This paper presents the phoneme recognizer for keyword spotting. Hidden Markov Model is used for modeling the phonemes. The keyword is described by phoneme transcription. Proposed method was examined at large speech corpus with various speakers. It is suggested to use such approach for speech information retrieval.

*Статья поступила в редакцию 30.06.2009.*